98 137
#2 US (CR)

# 日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。
This is to certify that the annexed is a true copy of the following application as filed with this Office.


出 願 年 月 日
Date of Application: １９９９年 ７月２３日

出 願 番 号
Application Number: 平成１１年特許願第２０９０９４号

出 願 人
Applicant (s): インターナショナル・ビジネス・マシーンズ・コーポレイション

１９９９年 ８月３１日

特 許 庁 長 官
Commissioner,
Patent Office 伴佐山建志

出証番号 出証特平１１－３０６１３３９

【書類名】　特許願

【整理番号】　JA998137

【提出日】　平成11年 7月23日

【あて先】　特許庁長官　伊佐山　建志　殿

【国際特許分類】　G06F 17/21

【発明の名称】　電子文書における文字情報の正規化方法

【請求項の数】　10

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名】　中居　治彦

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名】　木戸　彰夫

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名】　榎本　義彦

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名】　織田　哲治

【特許出願人】

【識別番号】　390009531

【氏名又は名称】　インターナショナル・ビジネス・マシーンズ・コーポレイション

【氏名又は名称原語表記】　INTERNATIONAL BUSINESS MACHINES CORPORATION

【代理人】

【識別番号】　100086243

【弁理士】

【氏名又は名称】　坂口　博

【代理人】

【識別番号】　100091568

【弁理士】

【氏名又は名称】　市位　嘉宏

【復代理人】

【識別番号】　100059258

【弁理士】

【氏名又は名称】　杉村　暁秀

【選任した復代理人】

【識別番号】　100072051

【弁理士】

【氏名又は名称】　杉村　興作

【選任した復代理人】

【識別番号】　100098383

【弁理士】

【氏名又は名称】　杉村　純子

【手数料の表示】

【予納台帳番号】　015093

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9304391

【包括委任状番号】　9304392

【プルーフの要否】　　要

【書類名】　明細書

【発明の名称】　電子文書における文字情報の正規化方法

【特許請求の範囲】

【請求項１】　電子文書中で使用されているフォントと、置換すべきターゲットフォントセット中のフォントとの比較を行うことにより、実際のフォント置換の際に参照されるフォント対照表を自動生成するフェーズと、自動生成されたフォント対照表を利用者に提示して、利用者が対照表の誤りを修正するフェーズと、修正されたフォント対照表を元に、電子文書中で実際にフォントの置換を行うフェーズと、からなることを特徴とする電子文書における文字情報の正規化方法。

【請求項２】　前記フォント対照表を自動生成するフェーズが、ソースとなる電子文書、その電子文書中で使用されているフォントセット、正規化を行うターゲットフォントセット、以前の変換で作製された対照表、字形比較の対象を限定するルールセットおよび漢字の部首ごとのマッピングに関するルールセットを記述したフォント対象情報を入力とし、フォント対照表の候補リストを出力する請求項１記載の電子文書における文字情報の正規化方法。

【請求項３】　似た文字間でのマッピングについて重み付け情報を参照ファイルとして出力する請求項２記載の電子文書における文字情報の正規化方法。

【請求項４】　前記フォント対照表の候補リストが、ソースフォント中の一文字と、それに対応する可能性があるターゲットフォント中の複数の文字との組を一要素とするリストである請求項２記載の電子文書における文字情報の正規化方法。

【請求項５】　前記ターゲットフォント中の複数の文字に対して優先順位情報を付加する請求項４記載の電子文書における文字情報の正規化方法。

【請求項６】　前記フォント対照表が、ソースフォントセットとそのソースフォントセット中の文字符号の組みと、ターゲットフォントセットとそのターゲットフォントセット中の文字符号の組みとの対応関係を要素とするリストである請求項１記載の電子文書における文字情報の正規化方法。

【請求項７】　前記フォント対照表を自動生成するフェーズにおけるフォントの比較を、ＯＣＲ（optical character reader）の技術を使用して自動的に行う請

求項１記載の電子文書における文字情報の正規化方法。

【請求項８】　前記フォント対照表の誤りを修正するフェーズが、フォント対照表の候補リストをエントリー毎に表示し、利用者にその候補の中から一つを選ばせる処理である請求項１記載の電子文書における文字情報の正規化方法。

【請求項９】　前記フォントの置換を行うフェーズが、フォント対照表と、ソースの電子文書の構造を記述したルールセットとを入力とし、ソース電子文書で使用されているフォントおよび文字符号の正規化を行う請求項１記載の電子文書における文字情報の正規化方法。

【請求項１０】　前記置換すべきフォントセットがユニコード・フォントによるフォントセットである請求項１記載の電子文書における文字情報の正規化方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、電子文書中の非標準のフォントセットを使用した文字を、対応する標準フォントセット中の文字で置き換えることにより、電子文書における文字情報の正規化を行う方法に関するものである。

【０００２】

【従来の技術】

従来、電子文書中のフォントの使用は、その文書の作成者に委ねられていた。また、ワープロ等の電子文書処理装置にインストールされているフォントは、機器毎に異なり、さらに、特定言語の基本的なものに限られていた。そのため、複数言語をある電子文書中で表現したいと思う作成者や、基本フォントセットに含まれていない文字を使用したいと思う作成者は、外字としてその文字のフォントを定義し、電子文書中で使用してきた。このことは、紙の上に印字された形での文書の交換に際しては問題に成り得なかったが、近年普及しつつある、インターネットを通じての電子的な文書の交換や、電子図書館への電子文書の登録に際しては、大きな問題となってきた。

【０００３】

電子文書の作成者と読者が、文字情報を正確に受け渡すためには、両者が同じ

フォントセットと文字符号を持たなければならない。しかし、プラットフォーム毎に使用できるフォントセットが異なるという現状を考えれば、情報交換に使われるフォーマット上、すなわち、インターネットの回線上を流れるフォーマット上や電子図書館もしくは企業内のセントラルファイル上に格納されるデータのフォーマット上では、文字情報は標準のフォントを用いた正規化されたものでなくてはならない。

【０００４】

【発明が解決しようとする課題】

従来より、電子文書の作成システム上ではフォントの置き換えは可能であったが、その置き換えは文字符号情報をそのままの形で保存して、フォント情報だけを他のフォントに置き換えるといったものであった。例えば外字フォントは、通常独立したフォントとして定義され、その中での文字のインデックスは文字の定義順に決められるのが普通であった。そのため、たとえユニコード・フォントのように世界中の主要な文字（通常の電子文書処理システムではサポートされていない数千語にわたるＪＩＳの補助漢字をも含む）の全てを含むような大きなフォントセットが使用されることになっても、フォント中の文字のインデックス（文字符号）が異なるために、フォントの置換を行うことができなかった。

【０００５】

あえてフォントの置換を行おうとすると、利用者は手動で電子文書中の文字符号の符号値を変えなければならなかった。また、それをするためには、利用者は元の電子文書中で使用されているフォントのインデックスと、置換先の対応する文字のインデックスを知らなければならない。電子図書館での電子文書の蓄積を考える場合、蓄積されるべき文書の作成者は不特定多数にのぼり、その全ての文書で使用されているフォントセットと、そのフォントセット中の文字のインデックスを記憶して、いちいち手動でフォントの正規化を行うことは実質的に無理であった。

【０００６】

結果として、従来の電子文書の文字情報を取り扱う電子図書館や社内のセントラルファイルにおいては、電子文書の正規化をあきらめ、作成されたままのかた

ちで文書の蓄積を行うしかなかった。そのため、電子文書の作成者と利用者のフォント環境の違いから文字化けが生じ、電子文書の交換に不都合が生じたり、Ｔｉｅｒ－０等の資源の限られたシステムにおいては、他のシステムで作成された電子文書を表示および処理できなかったりしていた。また、特開平７－３１９８５４号公報において、効果的な外字フォントファイルの作成と配布を目的とした外字管理システムが開示されているが、この技術は閉じられたネットワーク環境における外字フォントの管理に関するものであり、この技術をそのまま本発明の対象となる電子文書における文字情報の正規化に適用することはできなかった。

【０００７】

本発明の目的は上述した課題を解消して、プラットフォームもしくは電子文書作成システム毎に異なる様々なフォントを用いて作成された電子文書を、情報の質の劣化なくして、情報の蓄積および交換用にフォント使用の正規化を行うことができる電子文書における文字情報の正規化方法を提供しようとするものである。

【０００８】

【課題を解決するための手段】

本発明の電子文書における文字情報の正規化方法は、電子文書中の非標準のフォントセットを使用した文字を、対応する標準フォントセット中の文字で置き換えることにより、電子文書における文字情報の正規化を行う方法に関する。すなわち、電子文書中で使用されているフォントと、置換すべきターゲットフォントセット中のフォントとの比較を行うことにより、実際のフォント置換の際に参照されるフォント対照表を自動生成するフェーズと、自動生成されたフォント対照表を利用者に提示して、利用者が対照表の誤りを修正するフェーズと、修正されたフォント対照表を元に、電子文書中で実際にフォントの置換を行うフェーズと、から本発明の電子文書における文字情報の正規化方法を構成する。

【０００９】

本発明では、上述した構成をとることで、外字を使用して作られた電子文書の標準フォントセット例えばユニコード・フォントへの変換や部分的に外国語文書が存在する電子文書の標準フォントセットへの変換が可能となり、類似字形や外

国語文書の情報交換および蓄積が可能となる。

【００１０】

本発明の好適例として、フォント対照表を自動生成するフェーズが、ソースとなる電子文書、その電子文書中で使用されているフォントセット、正規化を行うターゲットフォントセット、以前の変換で作製された対照表、字形比較の対象を限定するルールセットおよび漢字の部首ごとのマッピングに関するルールセットを記述したフォント対象情報を入力とし、フォント対照表の候補リストを出力する。また、似た文字間でのマッピングについて重み付け情報を参照ファイルとして出力する。さらに、フォント対照表の候補リストが、ソースフォント中の一文字と、それに対応する可能性があるターゲットフォント中の複数の文字との組を一要素とするリストである。さらにまた、ターゲットフォント中の複数の文字に対して優先順位情報を付加する。また、フォント対照表が、ソースフォントセットとそのソースフォントセット中の文字符号の組みと、ターゲットフォントセットとそのターゲットフォントセット中の文字符号の組みとの対応関係を要素とするリストである。いずれの場合も、フォント対照表を自動生成するフェーズを好適に実施することができる。

【００１１】

また、本発明の好適例として、フォント対照表を自動生成するフェーズにおけるフォントの比較を、ＯＣＲ（optical character reader）の技術を使用して自動的に行う。さらに、フォント対照表の誤りを修正するフェーズが、フォント対照表の候補リストをエントリー毎に表示し、利用者にその候補の中から一つを選ばせる処理である。さらにまた、フォントの置換を行うフェーズが、フォント対照表と、ソースの電子文書の構造を記述したルールセットとを入力とし、ソース電子文書で使用されているフォントおよび文字符号の正規化を行う。また、置換すべきフォントセットがユニコード・フォントによるフォントセットである。いずれの場合も本発明を好適に実施することができる。

【００１２】

【発明の実施の形態】

図１は本発明の電子文書における文字情報の正規化方法の概念を説明するため

のフローチャートである。図１に従って本発明を説明すると、まず、電子文書中で使用されているフォントと、置換すべきフォントセット中の文字（フォント）との比較を行うことにより、実際のフォント置換の際に参照されるフォント対照表を自動生成するフォント対照表自動生成フェーズを実施し、フォント対照表の候補リストを作成する。次に、自動生成されたフォント対照表を利用者に提示して、利用者が対照表の誤りを修正するフォント対照表修正フェーズを実施し、新しいフォント対照表を作成する。最後に、修正されたフォント対照表を元に、電子文書中で実際にフォントの置換を行うフォント置換フェーズを実施し、正規化された電子文書を得ている。

【００１３】

上述した本発明の電子文書における文字情報の正規化方法を利用する可能性のある分野としては、電子図書館、文書管理システム、ＰＤＡ等のハンドヘルドデバイス（パベーシブコンピューティング環境）をサポートする中間サーバー、ＷＥＢパブリッシング、ＷＥＢブラウザー等があげられる。その一例として、利用者の作成した外字を含む電子文書をユニコード・フォントに正規化する場合を考える。この場合は、利用者が独自に定義した外字だけでなく通常の文字についてもユニコード・フォントに正規化する必要がある。通常の文字については、予め利用者の作成した電子文書のフォント例えばＭＳ明朝とユニコード・フォントとの間にフォントインデックスの対照表が存在するため、その対照表を元に簡単に正規化を行うことができる。

【００１４】

外字の正規化について、本発明の電子文書における文字情報の正規化方法を利用する。まず、各外字に対し、上述したフォント対照表自動作成フェーズを実施し、各外字に一致あるいは類似するユニコード・フォントを求め、フォント対照表の候補リストをフォント対照表として一旦作成する。通常、フォント対照表の候補リストは、各外字に対し複数のユニコード・フォントとなる。次に、フォント対照表修正フェーズを実施し、フォント対照表の候補リストを利用者に提示することで、利用者が対照表の誤りを修正、すなわち、候補リストの中から１つのフォントを選んだり、外字に対応するユニコード・フォントがない場合は、類似

するユニコード・フォントに割り付けたり、対応無しとしてユニコード・フォントの外字として登録したりして、フォント対照表の修正を行う。ユニコード・フォントは数千のＪＩＳ補助漢字をもサポートしているため、利用者の作成した外字のほとんどをユニコード・フォントに対応させることができる。最後に、修正されたフォント対照表を元に、電子文書中で実際にフォントの置換を行うフォント置換フェーズを実施し、ユニコード・フォントに正規化された電子文書を得ることができる。

【００１５】

以下、各フェーズ毎に詳細な説明を行う。

（１）フォント対照表自動生成フェーズについて：

本フェーズは、ソースとなる電子文書、その電子文書中で使用されているフォントセット、正規化を行うターゲットフォントセット、以前の変換で作成された対照表、字形比較の対象を限定するルールセットおよび漢字の部首ごとのマッピング（「一点しんにゅう」と「二点しんにゅう」、草冠の真ん中が切れているものと切れていないもの、などを同一の部首と認めるか認めないで別の文字とするか）に関するルールセットを記述したフォント対照指示情報を入力とし、フォント対照表の候補リストを出力する。この発明の好適な実装においては、本フェーズ実行において評価した似た文字間でのマッピングの重み付け情報を参照ファイルとして出力しておき、次回の実行の際に参照してもよい。

【００１６】

フォント対照表は、ソースフォントセットとそのフォントセット中の文字符号（フォントインデックス）の組と、ターゲットフォントセットとそのフォントセット中の対応する文字の文字符号の組との対応関係を要素とするリストである。フォント対照指示情報は、ソースのどのフォントセットをターゲットのどのフォントセットと対応ずけるか、および、ソースフォントセット中で字形比較の対照とするフォント群と、ターゲットフォント中の比較対照となるフォント群を指示する情報からなる。フォント対照表の候補リストは、ソースフォント中の一文字と、それに対応する可能性があるターゲットフォント中の複数文字との組を一要素とするリストである。本発明の好適な実装においては、ターゲットフォント中

の文字に対して優先順位情報を付加し、次のフェーズでの人手によるフォント対照表の確定作業の助けとすることもできる。

【００１７】

字形の比較はＯＣＲの技術を使用して以下の手順で行われる。

①ターゲットフォントセットから比較対照となる文字群の図形パターンを作成する。

②電子文書中の１文字を抜き出し、符号値を調べる。

③符号値が図形情報比較の対象となる文字のものであるなら、

Ａ．ソースフォントセットからその文字の図形パターンを作成する。

Ｂ．得られた図形パターンを、①で得られた文字群の図形パターンと比較し、似た図形パターンの組を対照表の候補リストに追加する。その際、好適な実装においては、候補リストの中での優先順位情報を追加する。

④上記②、③の処理を繰り返す。

【００１８】

（２）フォント対照表修正フェーズについて：

本フェーズは、前フェーズで得られたフォント対照表の候補リストと、以前の本フェーズの実行の結果得られたフォント対照表とを入力とし、最終的なフォント対照表を出力する。本フェーズでは、前フェーズで得られたフォント対照表の候補リストをエントリーごとに表示し、利用者にその候補の中から一つを選ばせることを主たる処理とする。本発明の好適な実装では、利用者が選んだ候補が、以前の処理で作成したフォント対照表のエントリーと矛盾が生じた場合、もしくは、多対一、一対多のマッピングを利用者が指示した場合、本フェーズの処理システムはその旨を利用者にワーニングとして指摘し、再考をうながしてもよい。また、本発明の好適な実装では、利用者にフォント対照表の候補リストを提示する際、候補となる文字の属性（文字の名前、文字の意味、文字種名、その他文字を特定するのに参考になる情報）を表示する機能を持たせてもよい。

【００１９】

（３）フォント置換フェーズについて：

前フェーズの出力であるフォント対照表と、ソースの電子文書の構造を記述し

たルールセットと、ターゲットの電子文書（ソースと同形式でもよい）の構造を記述したルールセットとを入力とし、ソース電子文書で使用されているフォントおよび文字符号の正規化を行う。この際、ソースとターゲット間で電子文書の形式および構造が異なった場合、このフェーズで電子文書の形式変換を同時におこなってもよい。

【００２０】

以上詳細に説明した本発明の電子文書における文字情報の正規化方法は、上述した本発明の目的である、プラットフォームもしくは電子文書作成システムごとに異なる様々なフォントを用いて作成された電子文書を、情報の質の劣化なくして、情報の蓄積および交換用にフォント使用の正規化を行うこと、に加えて、以下に列記する様々な目的に使用でき様々な効果を得ることができる。

(1) 様々な文字符号を用いて作成された電子文書の文字符号を、インターネット標準の多国悟文字符号であるユニコードに変換することにより、インターネット標準の文書記述言語であるＸＭＬへの変換を容易にし、電子文書のＷＥＢ上での公開を可能にすることができる。

(2) 様々な文字符号を用いて作成された電子文書の文字符号を、特定の文字符号に正規化することにより、電子図書館や企業のセントラルファイル上での情報検索の効率を上げることができる。

【００２１】

(3) 正規化された文字符号、文字フォントを用いて蓄積および交換される電子文書を、クライアント環境においてクライアント環境特有の文字符号およびフォントに逆変換することにより、情報交換に用いられるフォントを持たない環境でも、類似の文字を使用してその電子文書を表示することを可能にすることができる。

(4) 正規化された文字符号、文字フォントを用いて蓄積および交換される電子文書を、クライアント環境においてクライアント環境特有の文字符号およびフォントに逆変換することにより、クライアント環境での情報処理の効率をあげることができる。

(5) 文字符号およびフォントの正規化に際して参照される対照表、もしくは、置

換のための命令セットを半自動的に行うことにより、利用者の付加を減らし、前記の文書の正規化のための作業量を実質的な範囲まで引き下げることができる。

【００２２】

(6) Ｔｉｅｒ－０等の資源の少ない（フォントが少ない、もしくは、効率的な文書処理のために、元の電子文書の文字符号からそのシステムの文字符号への変換テーブルおよび機能を持たない）システムで、他のシステムで作成された文書の表示および処理をさせようとした場合、その電子文書のアクセスの際アクセス経路上の中間サーバーに対して、その資源の少ないクライアント環境にオプティマイズした文書の正規化を依頼することにより、クライアント環境で処理可能な形式に電子文書の形式変更を行わせることができる。

(7) 従来単独でしか行えなかったフォントや文字符号の置換を、過去において行われた別の文書の正規化を参照しながら行うことによって、文字符号およびフォントの対照表の自動生成の効率を上げることができる。

(8) 過去において行われた別の文書の正規化を参照しながら文字符号およびフォントの対照表の自動生成を行うことによって、過去のマッピングの経験を生かし、マッピングミスの可能性を削減することができる。ここで、マッピングミスとしては、以下の場合が考えられる。

・ターゲットのフォントセットの中に、字形の似た文字が複数存在した場合、間違った方のマッピングをとってしまうこと、もしくは、マッピングにゆれが生じてしまうこと。

・ソースのフォントセットの中に、字形の似た文字が複数存在した場合、その複数の文字をターゲットの一つの文字にマップしてしまうこと。

【００２３】

(9) ソースとターゲットもフォントセットの中で、自動的に字形の比較を行うフォント群を特定することによって、マッピングが事前に定義できる文字（フォント）についての比較を避け、対照表自動生成の効率を上げることができる。

(10) ソースとターゲットもフォントセットの中で、自動的に字形の比較を行うフォント群を特定することによって、マッピングが事前に定義できる文字（フ

ォント）についての比較を避け、利用者の意図しない対照表が生成される（例えば、ＪＩＳの第一水準の文字がＪＩＳの第二水準の文字にマップされる）危険性を低減することができる。

(11) ソースに含まれる特定のフォントを比較するターゲットのフォントを規定することによって、タイプフェースが異なるフォントの比較によるマッピングの正確性の低下を低減することができる。

(12) 言語とその言語によって使用されるフォントセットとの関係に注目することにより、対照表自動生成時においてリンギスティックなルールの導入を可能にすることができる。このことにより、字形情報の比較によって得られた変換対象文字を、前後の文字と接続して単語にし、さらに対照表の自動生成の精度を上げることができる。

【００２４】

【発明の効果】

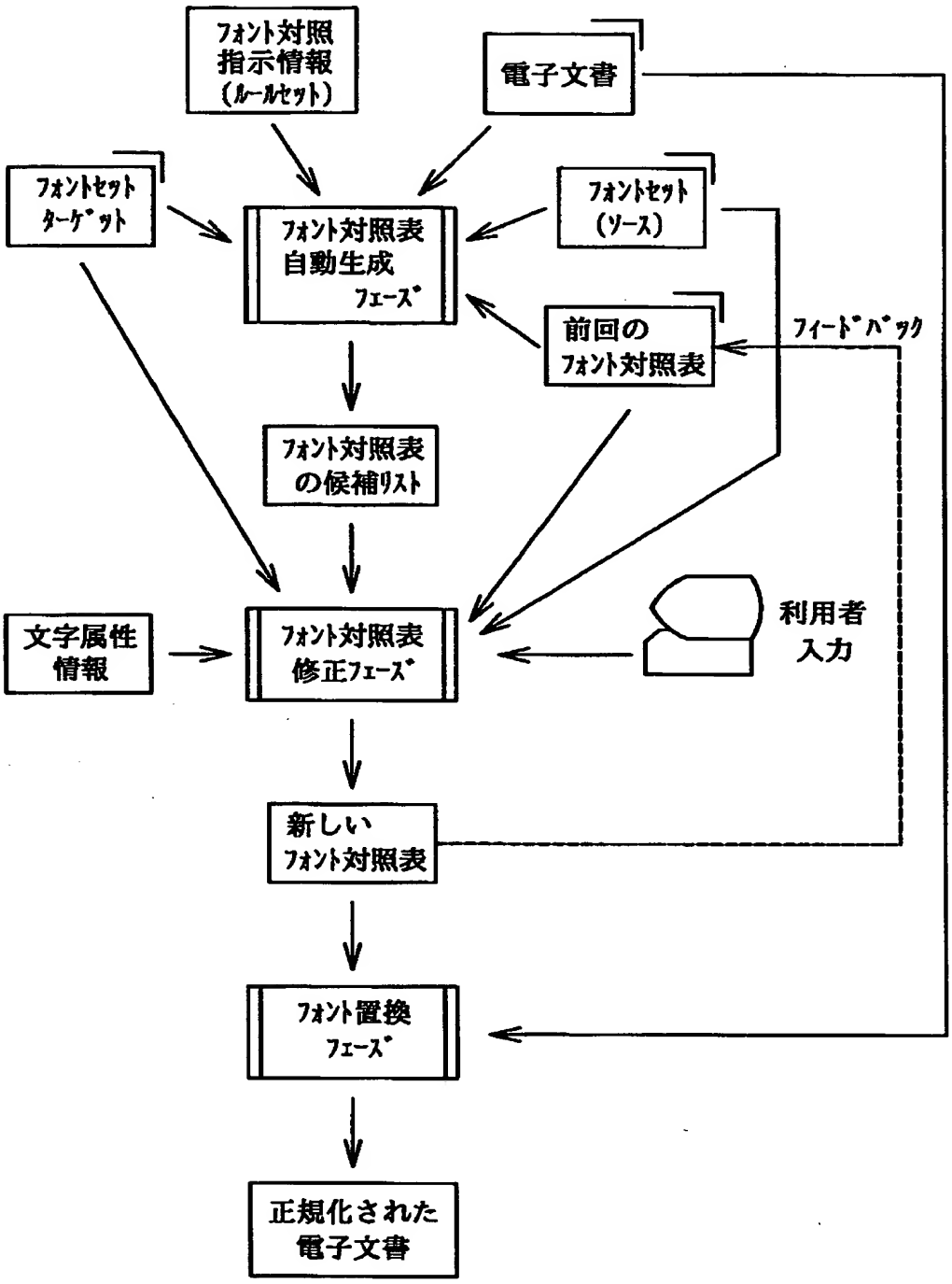
以上の説明から明らかなように、本発明によれば、外字を使用して作られた電子文書の標準フォントセット例えばユニコード・フォントへの変換や部分的に外国語文書が存在する電子文書の標準フォントセットへの変換が可能となり、類似字形や外国語文書の情報交換および蓄積が可能となる。

【図面の簡単な説明】

【図１】　本発明の電子文書における文字情報の正規化方法の概念を説明するためのフローチャートである。

【書類名】　　図面

【図１】


フォント対照
指示情報
(ルールセット)
電子文書
フォントセット
ターゲット
フォントセット
(ソース)
フォント対照表
自動生成
フェーズ
前回の
フォント対照表
フィードバック
フォント対照表
の候補リスト
文字属性
情報
フォント対照表
修正フェーズ
利用者
入力
新しい
フォント対照表
フォント置換
フェーズ
正規化された
電子文書

【書類名】　　要約書

【要約】

【課題】　プラットフォームもしくは電子文書作成システム毎に異なる様々なフォントを用いて作成された電子文書を、情報の質の劣化なくして、情報の蓄積および交換用にフォント使用の正規化を行うことができる電子文書における文字情報の正規化方法を提供する。

【解決手段】　電子文書中で使用されているフォントと、置換すべきターゲットフォントセット中のフォントとの比較を行うことにより、実際のフォント置換の際に参照されるフォント対照表を自動生成するフェーズと、自動生成されたフォント対照表を利用者に提示して、利用者が対照表の誤りを修正するフェーズと、修正されたフォント対照表を元に、電子文書中で実際にフォントの置換を行うフェーズと、から本発明の電子文書における文字情報の正規化方法を構成する。

【選択図】　　図１

# 出　願　人　履　歴　情　報

識別番号　　〔３９０００９５３１〕

１．変更年月日　　１９９０年１０月２４日
［変更理由］　新規登録
住　所　アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　（番地なし）
氏　名　インターナショナル・ビジネス・マシーンズ・コーポレイション